UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

GENERATING REAL-WORLD AND SYNTHETIC GRAPH DATASET FOR GNN APPLICATION

Part 2

21st July 2022

The IMPACT Research Group

CHALLENGES TO GRAPH NEURAL NET COMMUNITY

Dataset	Date	Node emb size	#nodes (millions)	#edges (millions)	#labeled nodes (millions)
OBGN-papers	2020	128*	111	1,615	1.4
MAG240M	2021	768	260	1,300	1.4
PinSAGE dataset	2018	128-2K (avg. 1K)	3000	18,000	UNK

- Large gap between publicly available graph datasets and ones used by industry.
- Our goal is to propose a new dataset that will help both system designers and GNN researchers in two ways:
  - Given a dataset schema, propose a methodology to generate arbitrary sized graphs (homogenous or heterogenous) and node embeddings with prescribed number of nodes, edges and relations.
  - Provide a dataset with synthetic node embeddings for system developers and another dataset with node embeddings generated using NLP methods for GNN researchers and system developers.

* → paper does not provide information. Calculated via dataset size.

Creating a graph dataset is not an easy task

- Finding large open-source real-world dataset that has multiple relations and properties is quite a challenge
- Large number of resources required to handle huge quantities of raw data and cleaning it.
- Generating labels is not trivial -> requires humans to label.
- Single databases may not have all the required information and can result in poor graph generation-> merge databases.

How to make the dataset useful?

We want to study the impact on the creation of a large dataset since it hasn't been done in the community

- Impact of dataset size.
- Impact of labelled data.
- Impact of node embedding - creation model, size, language

Dataset	Date	Node emb size	#nodes (millions)	#edges (millions)	#labeled nodes (millions)
OBGN-papers	2020	128*	111	1,615	1.4
MAG240M	2021	768	260	1,300	1.4

We do not want to generate a graph dataset naively.

METHODOLOGY TO CREATE A GRAPH DATASET

1. Database – We need real world data in order to create a real-world graph dataset.

We start by looking into the Microsoft Academic Graph – MAG – which is an open-source database.

MICROSOFT ACADEMIC GRAPH (MAG) DATABASE

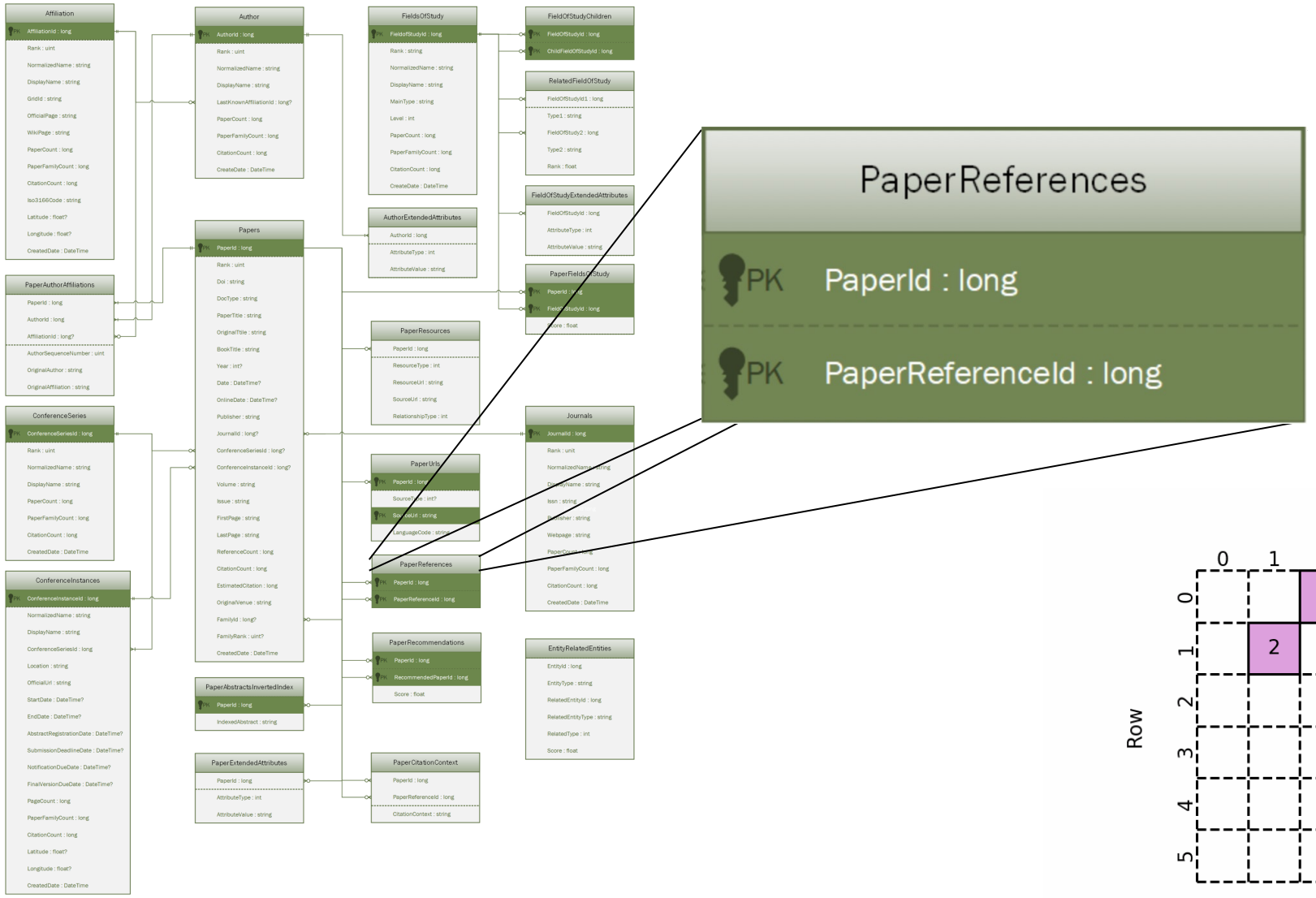
266,988,075

Unique papers in the MAG database

7.11

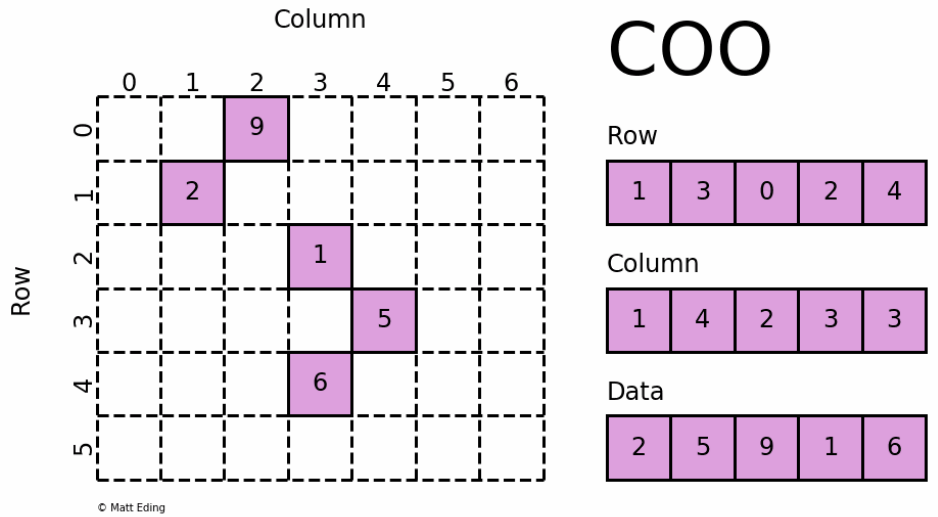
Average citation per paper

MICROSOFT ACADEMIC GRAPH (MAG) DATABASE



1,900,288,267

Paper paper edges in MAG



COO

Row

1	3	0	2	4

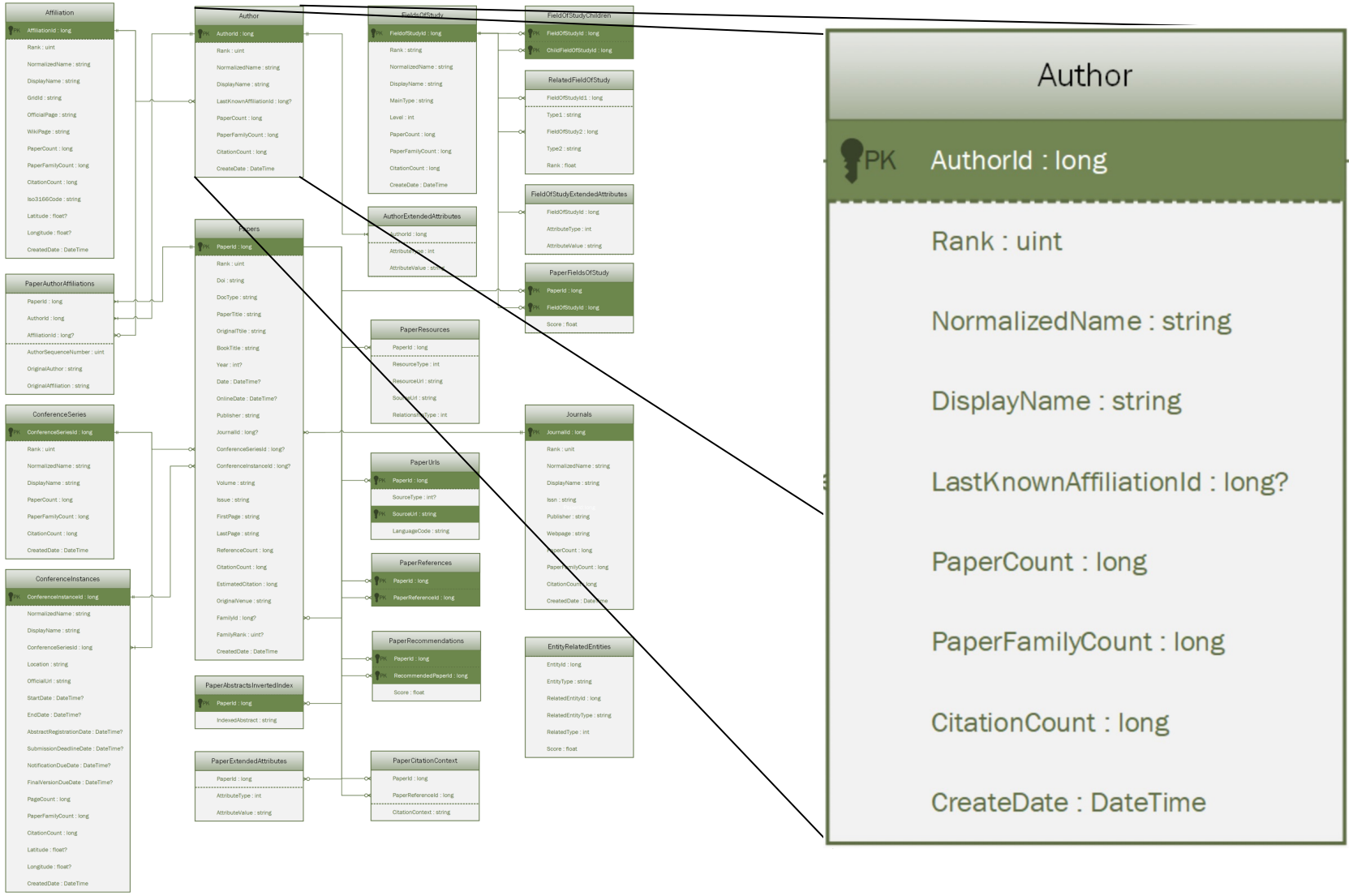
Column

1	4	2	3	3

Data

2	5	9	1	6

MICROSOFT ACADEMIC GRAPH (MAG) DATABASE



279,356,515

Unique authors in the MAG database

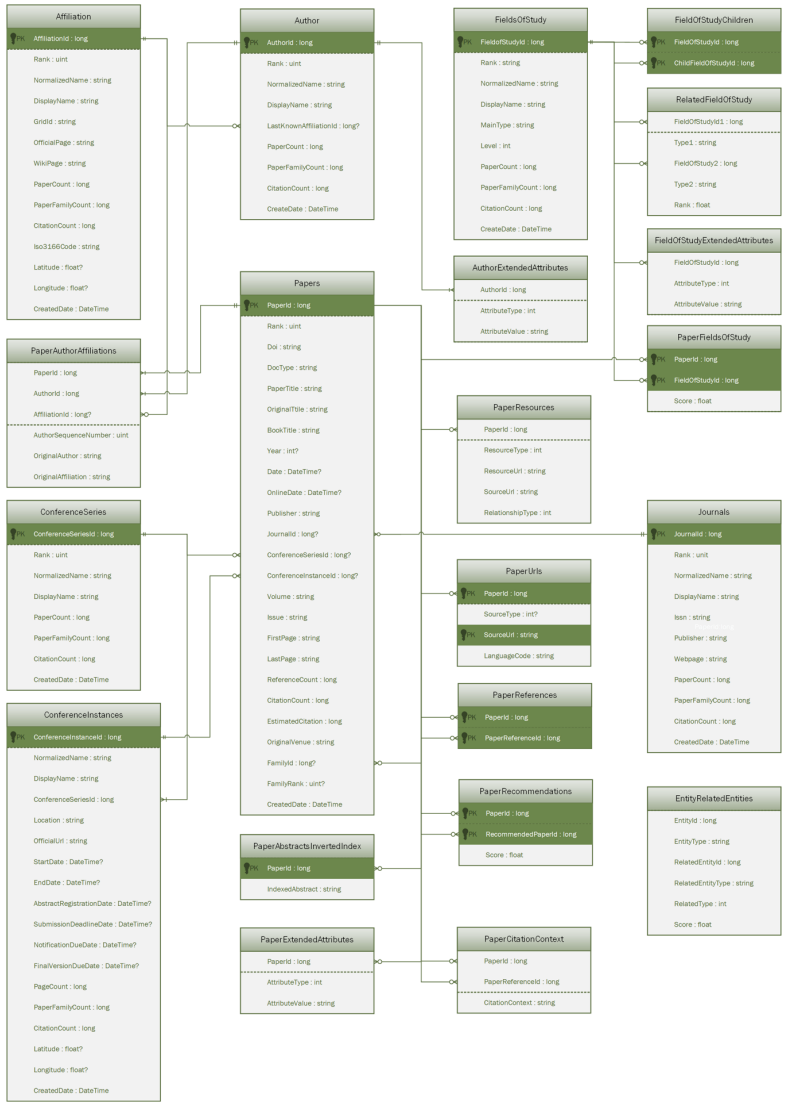
2.71

Average authors per paper

2.59

Average paper per author

MICROSOFT ACADEMIC GRAPH (MAG) DATABASE



94,448,476

Papers with journal IDs

266,988,075

Papers with titles/abstract

49,053

Number of unique journals

100%

% of Papers with title/abstract

5,152,118

Papers with conference IDs

35.37%

% of Papers with journal

4548

Number of unique conferences

1.93%

% of Papers with conference

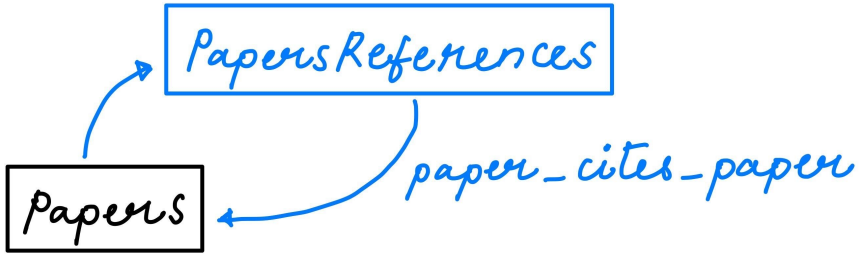
METHODOLOGY TO CREATE A GRAPH DATASET

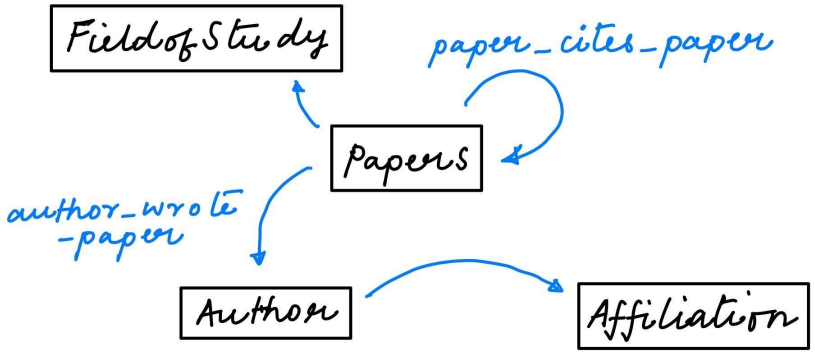
1. Database – We need real world data in order to create a real-world graph dataset.

We start by looking into the Microsoft Academic Graph – MAG – which is an open-source database.

2. Graph Dataset Schema – What are the nodes and edges in the dataset? Is it homogeneous or heterogenous?

IGB260 is a homogeneous graph database and IGBH600 is a heterogeneous graph dataset.





POPULAR GRAPH DATASETS METRICS (HOMOGENEOUS)

Dataset	#nodes (millions)	# nodes labelled (millions)	#edges (millions)	Node Degree (max/avg/min)	Node emb size	Homophily
ogbn-arxiv	0.2	0.2	1	*/5/1	128	65%
ogbn-mag	0.7	0.7	11	*/15.7/1	768	48%
ogbn-papers100M	111	1.4	1,615	*/14.5/1	768	*
MAG240M (Largest Available)	122	1.4	1,300	*/10.67/1	768	*
IGB-tiny (exp)	0.008	0.008	0.02	51 / 2.56 / 1	364 - 1024	52%
IGB-small	0.2	0.2	0.2	* / 1 / 1	364 - 1024	54%
IGB-medium	10	10	26	* / 2.6 / 1	364 - 1024	62%
IGB-large	100	100	700*	* / 7 / 1	364 - 1024	67%*
IGB260M	267	150	2,000*	273,370 / 7.49* / 1	364 - 1024	69.5%

HOMOPHILY:

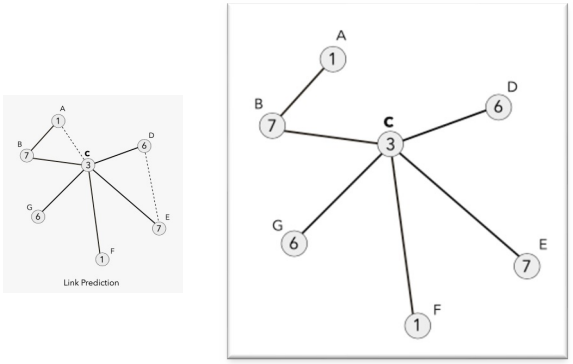
$$\beta = \frac{1}{|V|} \sum_{v \in V} \frac{\text{Number of } v\text{'s neighbors who have the same label as } v}{\text{Number of } v\text{'s neighbors}}$$

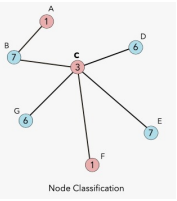
* Work in progress

METHODOLOGY TO CREATE A GRAPH DATASET

3. Decide the downstream task – node level, edge level or graph level.

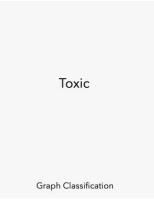
IGB is primarily meant to be a multi label node classification task dataset however, it can be modified to be an edge prediction task dataset as well.





4. Find ground truth labels for your data according to your downstream task.

For node level you need node labels, for edge prediction its easier since the graph edges can be masked and used as train/test data. In graph level you need graph labels.



ARXIV LABELS

In every graph database based on MAG dataset with a subject based node classification the arXiv dataset was used for getting labels.

1,398,159

Papers have arXiv labels

0.52%

Papers have arXiv label

172

Total distinct fields of study

arXiv
Category Taxonomy

- Computer Science
- Economics
- Electrical Engineering and Systems Science
- Mathematics
- Physics
- Quantitative Biology
- Quantitative Finance
- Statistics

Pros:

- 1.4 million labels
- 172 classes structured into 8 subjects

Cons:

- Covers 0.5% of MAG paper nodes.
- Can't judge whether model isn't getting enough data.

SEARCH FOR BETTER LABELS – SEMANTIC SCHOLAR

Semantic **Scholar**

'Art', 'Biology', 'Business', 'Chemistry', 'Computer Science', 'Economics', 'Engineering', 'Environmental Science', 'Geography', 'Geology', 'History', 'Material Science', 'Mathematics', 'Medicine', 'Philosophy', 'Physics', 'Political Science', 'Psychology', 'Sociology'

Pros:

- Firstly, majority of the graph is now labelled. (x100 more than before)
- Secondly, this gives us even more nodes.
- Provides 19 distinct subject classes.

157,675,969

Papers have a Semantic Scholar (SS)

59%

Papers have a SS label

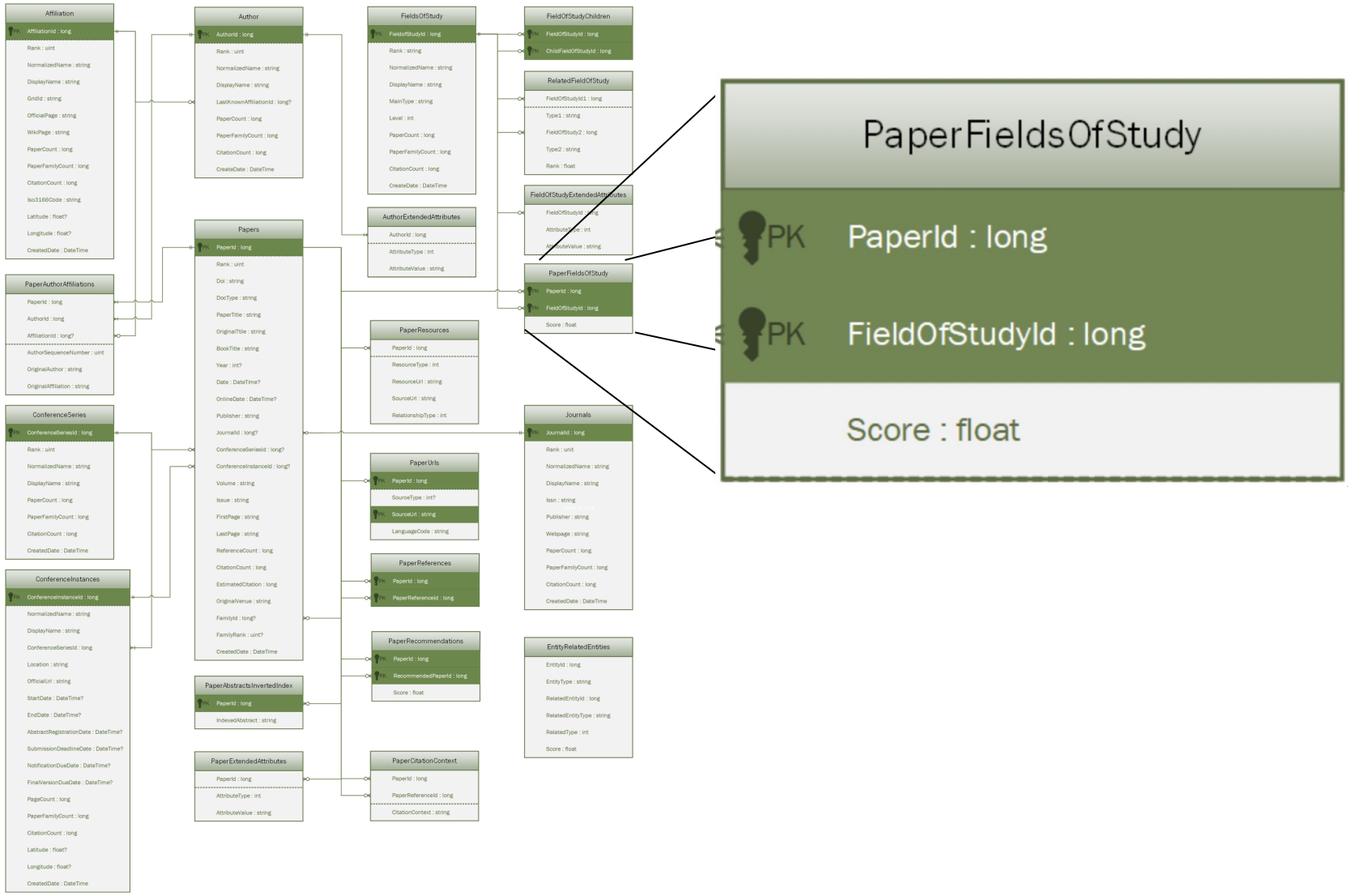
19

Total distinct fields of study

2.2 B

Number of paper citation edges

FIELD OF STUDY LABELS AS POTENTIAL ALTERNATIVE LABEL



207,677,865

Papers have a Field of Study (FoS)

78%

Papers have a FoS label

714,373

Total distinct fields of study

#1: DO MORE LABELS HELP IMPROVE ACCURACY?

HYPOTHESIS: We expect GNN models to have better accuracy since these are supervised learning models.

OPEN QUESTIONS:

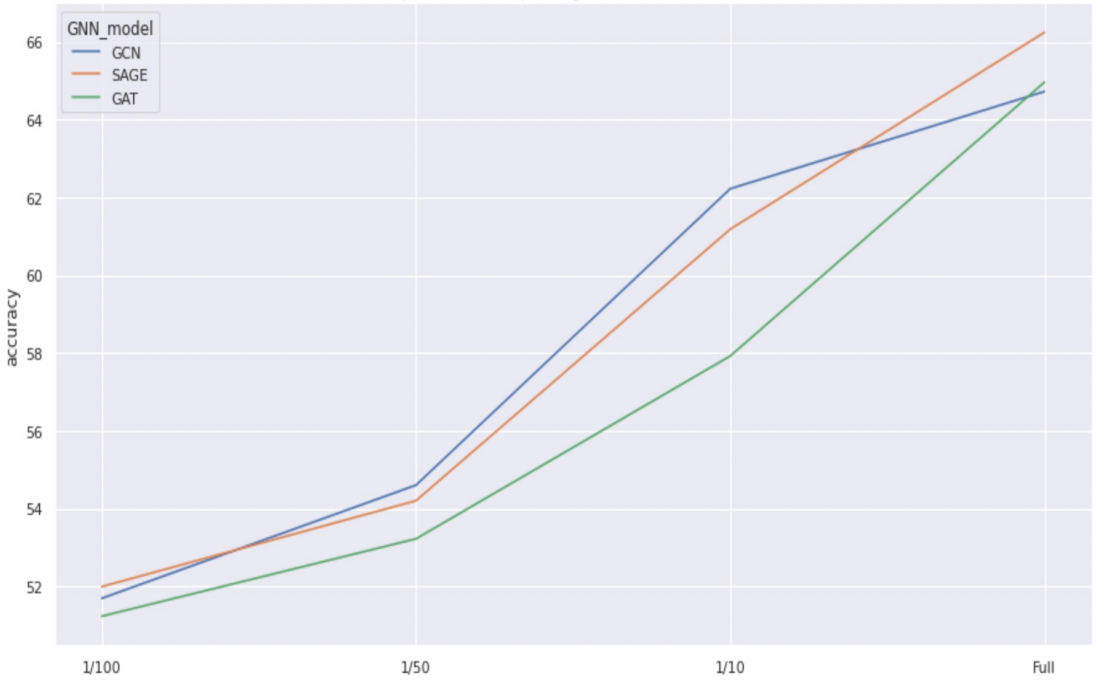
- What is the reason for only a 15% gain in performance with 100x more labelled data ?
- Is there a point where accuracy v data plateaus?

EXPERIMENT (batched-GraphSAGE):

GRAPH	EDGES	NODES	FULL LABEL	1/100** LABEL	DIFFERENCE
IGB-tiny	8,232	21,060	64.78%	49.75%	15.03
IGB-small	168,319	168,323	67.38%	51.45%	15.93
IGB-med	10,000,384	26,517,372	68.79%	56.28%	12.51

** Our dataset has 100x more labelled data than MAG240M

GNN performance based on labelled data on IGB-small

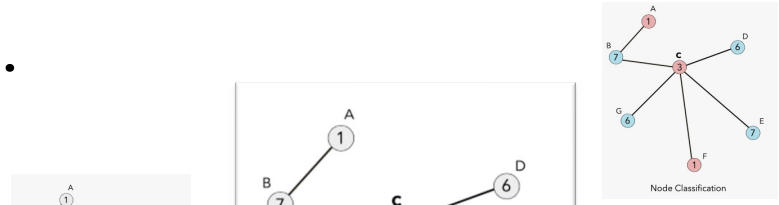


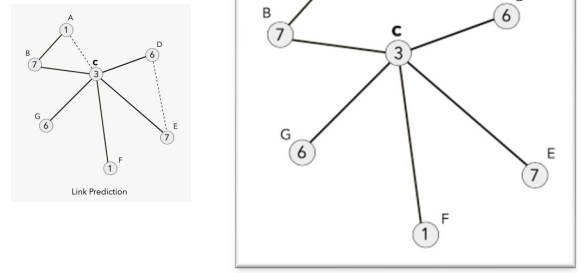
Fraction of labelled data

CREATING NODE EMBEDDINGS

3. Decide the downstream task – node level, edge level or graph level.

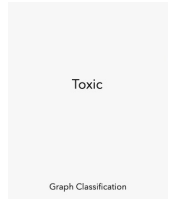
IGB is primarily meant to be a multi label node classification task dataset however, it can be modified to be an edge prediction task dataset as well.





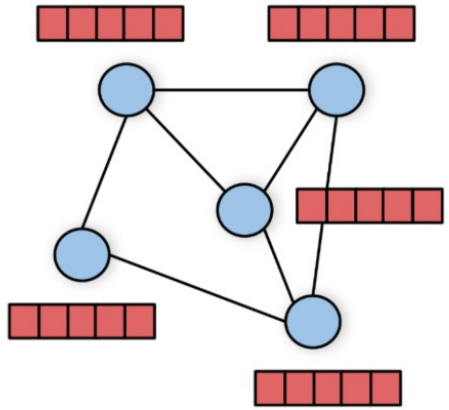
4. Find ground truth labels for your data according to your downstream task.

For node level you need node labels, for edge prediction its easier since the graph edges can be masked and used as train/test data. In graph level you need graph labels.



5. Initialize the graph nodes with a n-dim embedding.

Graphs need node embeddings as these become the input to the GNN model.

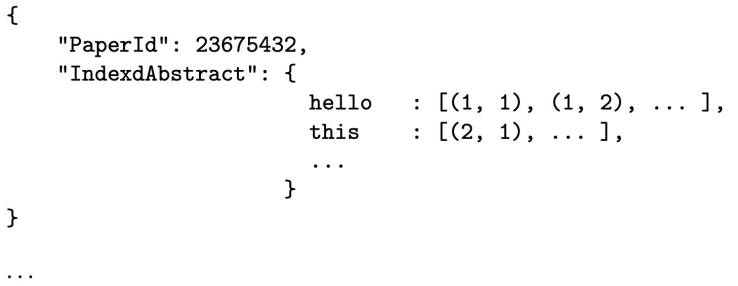


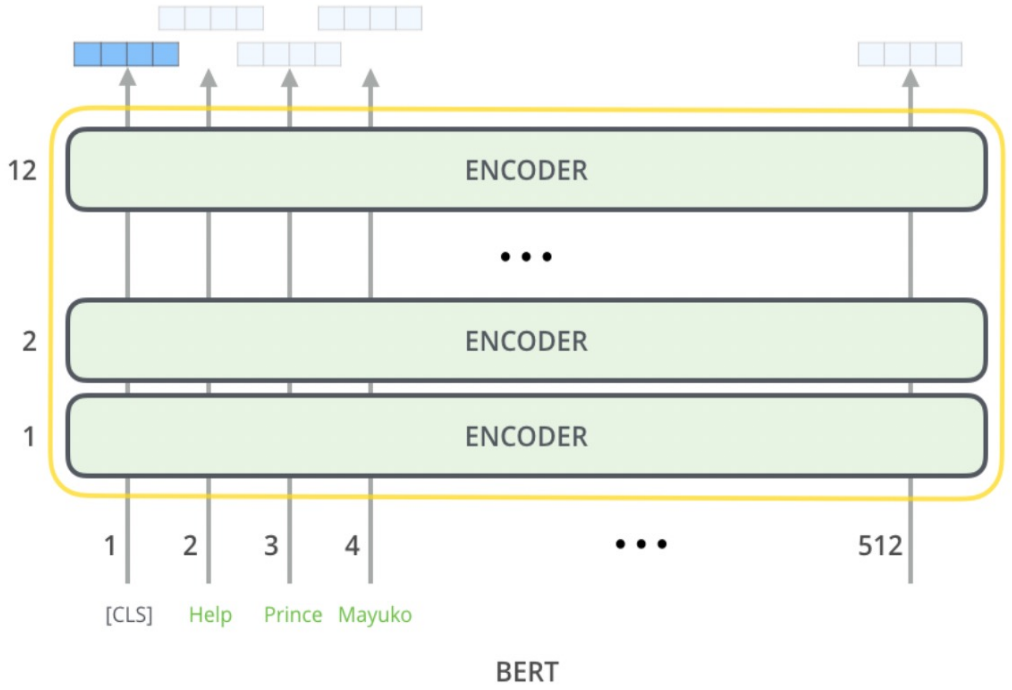
GENERATION OF NODE EMBEDDINGS

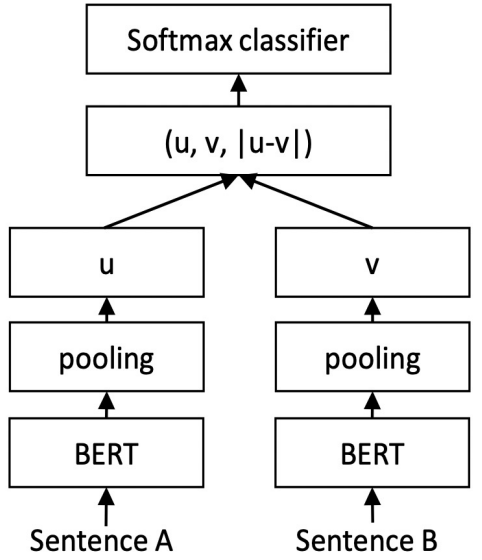
From the MAG dataset we extracted abstracts for each paper which is stored in an inverted index format.

After converting it into text and cleaning the data, we used the Sentence-BERT model to generate node embeddings for each paper node.

The BERT model takes the entire abstract and tokenizes each word using a layer of Encoders which is then pooled together to get a single embedding using the Sentence-BERT model.





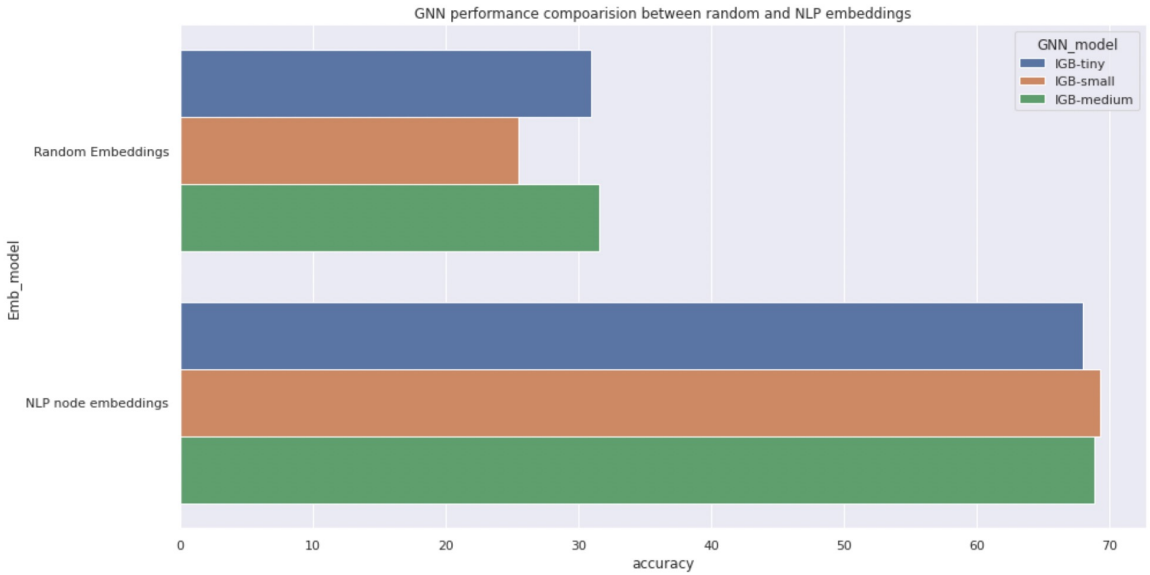


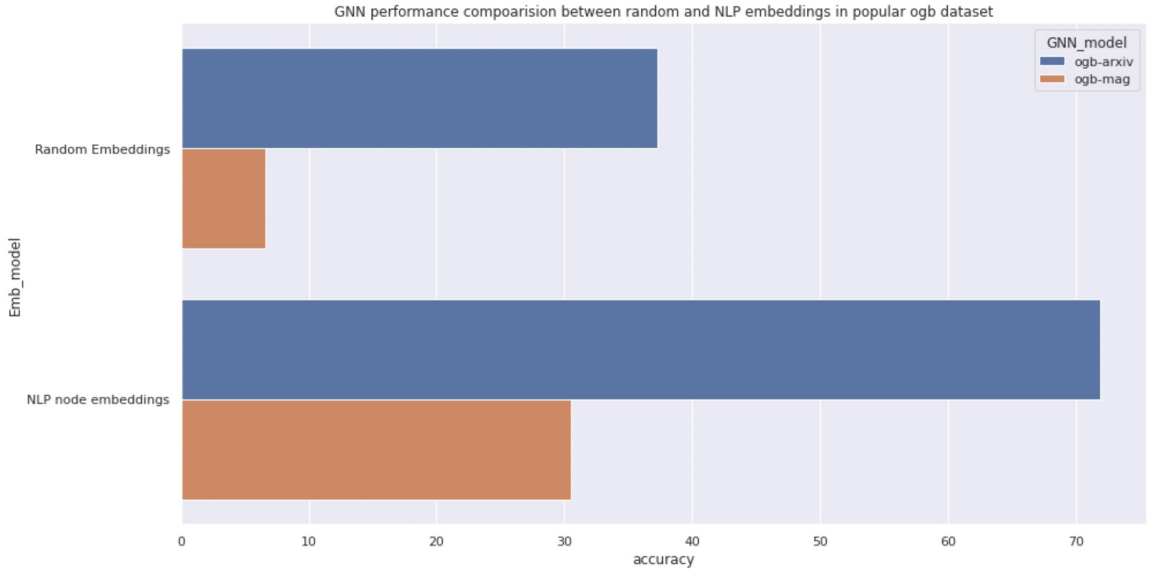
Alammar, J (2018). The Illustrated Transformer [Blog post]. Retrieved from https://jalammar.github.io/illustrated-transformer/

SYNTHETIC RANDOM NODE EMBEDDINGS IS BAD IDEA FOR MODEL!

HYPOTHESIS: We expect a X% increase in GNN model accuracy if we initialize the nodes with embeddings generated by sentence transformers based on the paper title and abstract instead of random initialization.

- **>50% drop in performance**
- **node embeddings gives the GNN model context thus leading to much higher accuracy**.





*embeddings created with roberta

INFLUENCE OF NLP MODELS AND PROPERTIES ON GNN

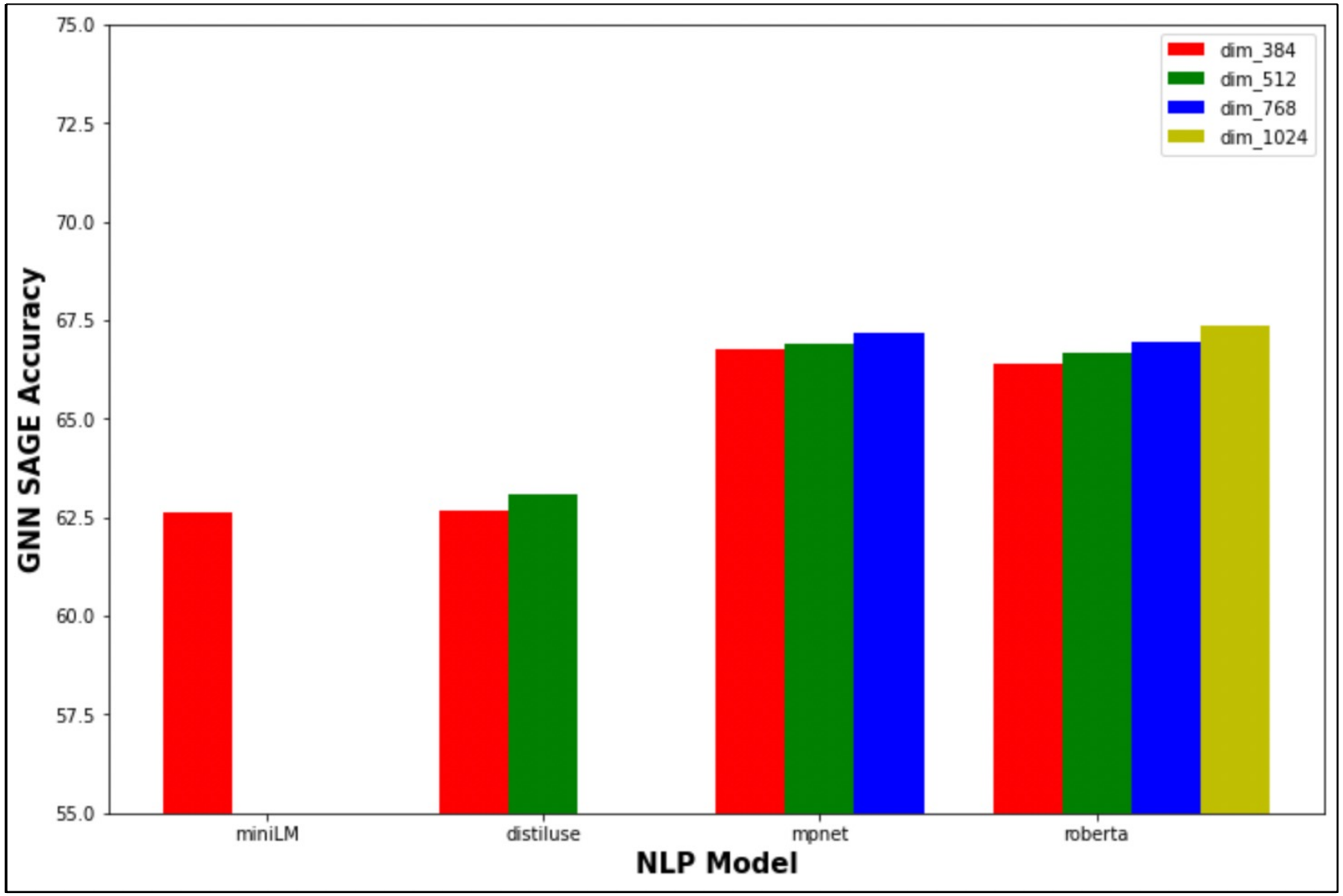
1. Does the NLP model matter?
2. What about embedding size?
3. Does NLP model accuracy matter?
4. Does language in which NLP model is trained matter?

NLP MODEL	EMB SIZE	AVG. PERF	MODEL SIZE
all-MiniLM-L6-v2	384	58.80	80 MB
distiluse-base-multilingual-cased-v1	512	45.59	480 MB
all-mpnet-base-v2	768	63.30	420 MB
all-roberta-large-v1	1,024	61.64	1360 MB

Hugging Face

All data taken from Hugging Face model cards -> https://www.sbert.net/docs/pretrained_models.html#sentence-embedding-models/

NODE EMBEDDINGS EXPERIMENTS



How does the NLP model impact performance?

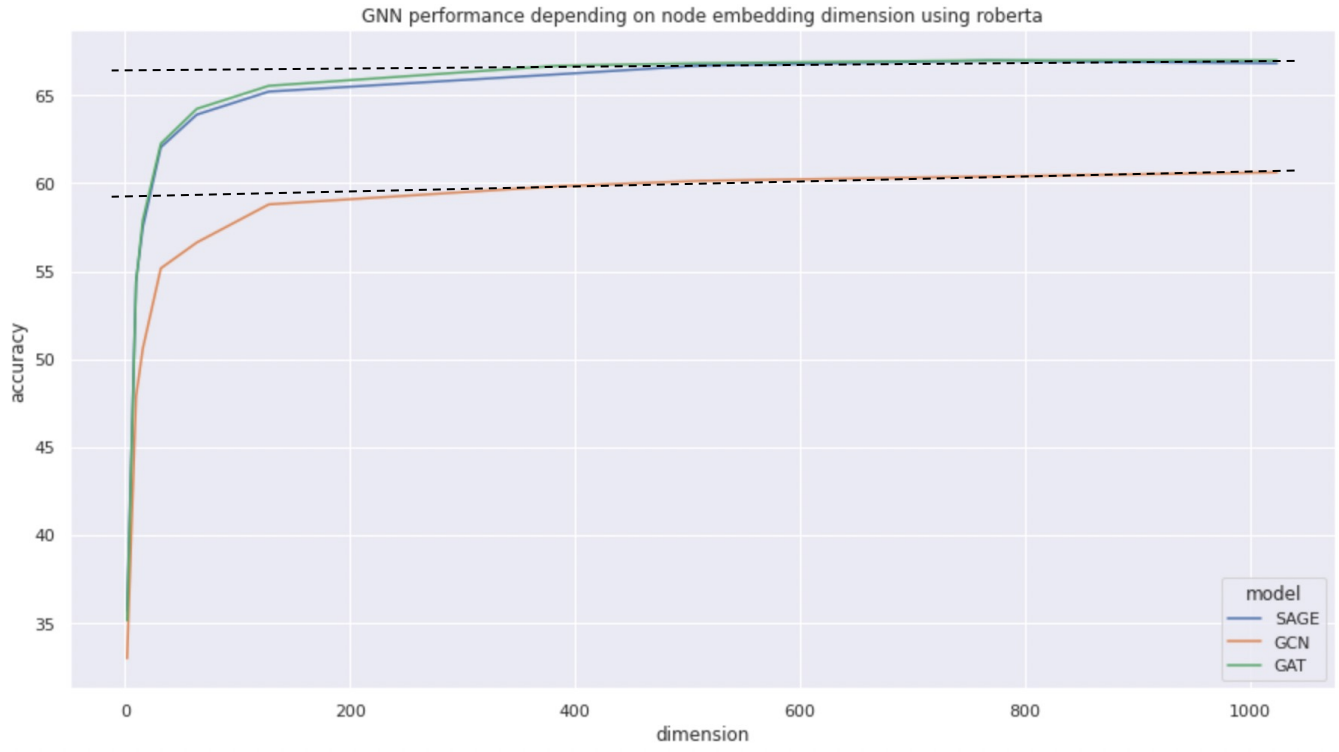
As we can see, the NLP model that generates embedding matters.

DOES EMBEDDING SIZE MATTER AND WHAT IS THE MINIMUM SIZE?

In order to test this, we first generated Roberta embeddings and then used PCA analysis to reduce the dimensions to
768, 512, 364, 256, 128, 64, 32, 16, 8, 2

GNNs appear to have a strong tolerance for node embedding sizes as you can see reducing a 1024 dim node embedding to ~300 didn't cause a significant drop. However, from 200 to 2 the model performance deteriorates greatly.

This experiment shows that embedding size matters for GNN model performance, but we can do dimensionality reduction to save storage without affecting the performance greatly.



LANGUAGES PRESENT IN MAG DATASET FOR PAPERS

- 'Japanese': 14,561,007,
- 'Spanish; Castilian': 7,120,422,
- 'Chinese Simplified': 5,914,056,
- 'Korean': 4,685,949,
- 'French': 4,075,241,
- 'German': 2,632,945,
- 'Russian': 2,123,281,
- 'Portuguese': 2,017,665,
- 'Indonesian': 1,673,950,
- 'Polish': 1,089,705,
- 'Italian': 1,031,295,

>80

Languages of paper in MAG

11

Languages have > 1M nodes

EXTRACTING EMBEDDING WITH LANGUAGE SPECIFIC MODEL

QUESTION: How does language affect the node embeddings and GNN?

- Created 3 graphs with paper nodes of only that language.
- No stat-sig impact on GNN performance even though the NLP model itself gives has different performance based on language

Hypothesis: Here the GNN overcomes the challenge faced by the NLP, since it learns from the structure of the graph which makes it language agnostic.

Japanese - GAT

Graph(num_nodes=1,975,296, num_edges=3,291,665)

- 384 eng -> 69.32288766263352 (Trained on english)
- 384 all -> 68.340252113603 (Same model trained on 50 languages)
- 512_v1 -> 67.09335290841898 (Not trained on japanese)
- 512_v2 -> 67.44367944109756 (Model trained on 50 languages including japanese)

Spanish - GAT

Graph(num_nodes=1,207,296, num_edges=1,266,183)

- 512_v1 -> 57.48736850824153 (Model trained on spanish)
- 512_v2 -> 57.04712995941357 (Model trained on 50 languages including spanish)

French - GAT

Graph(num_nodes=841,728, num_edges=2*844,101)

- 384 eng -> 60.390740553737224 (Trained on english)
- 384 all -> 59.4830914717815 (Same model trained on 50 languages)
- 768 eng -> 60.15848218263468 (Trained on english)
- 768 all -> 60.251741937783265 (Same model trained on 50 languages)

IGB REAL WORLD GRAPH DATASETS

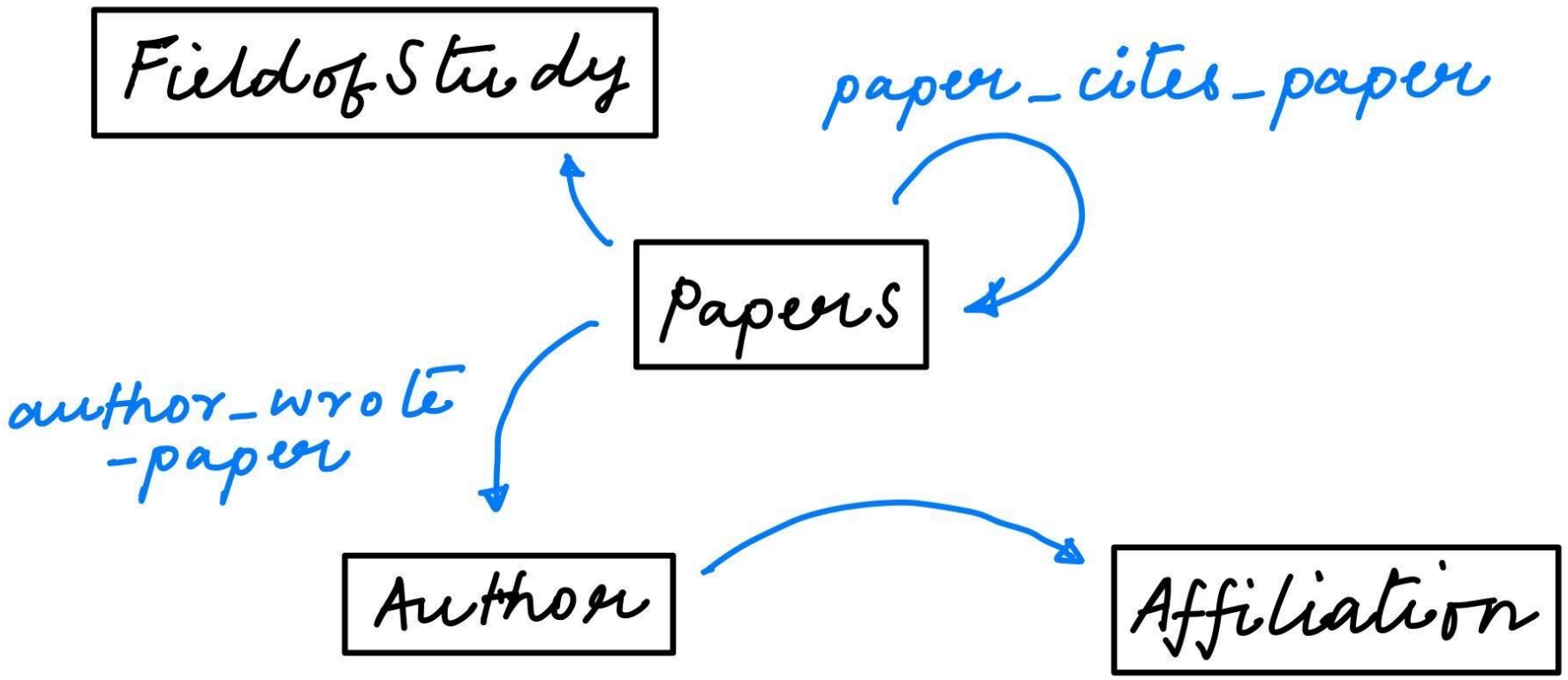
Dataset	#nodes (millions)	# labelled (millions)	#edges (millions)	Node Degree (max/avg/min)	Node emb size	Homophily
ogbn-arxiv	0.2	0.2	1	*/5/1	128	65%
ogbn-mag	0.7	0.7	11	*/15.7/1	768	48%
ogbn-papers100M	111	1.4	1,615	*/14.5/1	768	*
MAG240M (Largest Available)	122	1.4	1,300	*/10.67/1	768	*
IGB-tiny (exp)	0.008	0.008	0.02	51 / 2.56 / 1	364 - 1024	52%
IGB-small	0.2	0.2	0.2	* / 1 / 1	364 - 1024	54%
IGB-medium	10	10	26	* / 2.6 / 1	364 - 1024	62%
IGB-large	100	100	700*	* / 7 / 1	364 - 1024	67%*
IGB260M	267	150	2,000	273,370 / 28.32 / 1	364 - 1024	69.5%

Dataset	Date	Type	Node emb size	#nodes (millions)	#edges (millions)	#labeled nodes (millions)
OBGN-papers	2020	Real	128+	111	1,615	1.4
MAG240M	2021	Real	768	260	1,300	1.4
**IGB-260M	2022	Real/Synthetic	128 - 4kB	267	1,900	~150+
**IGBH-600M	2022	Real/Synthetic	128 - 4kB	~600+	~ 3,000+	~400+
PinSAGE dataset (proprietary dataset)	2018	Real	128-2K	3000	18,000	NA

* Work in progress

+Speculated

IGBH600M -> HETEROGENEOUS GRAPH



NEXT STEPS

- Run the same tests on the heterogeneous graph.
- Create a multirelational graph
- Get the metrics for both

Thank you!

Please feel free to ask me any questions

The Grainger College of Engineering

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN